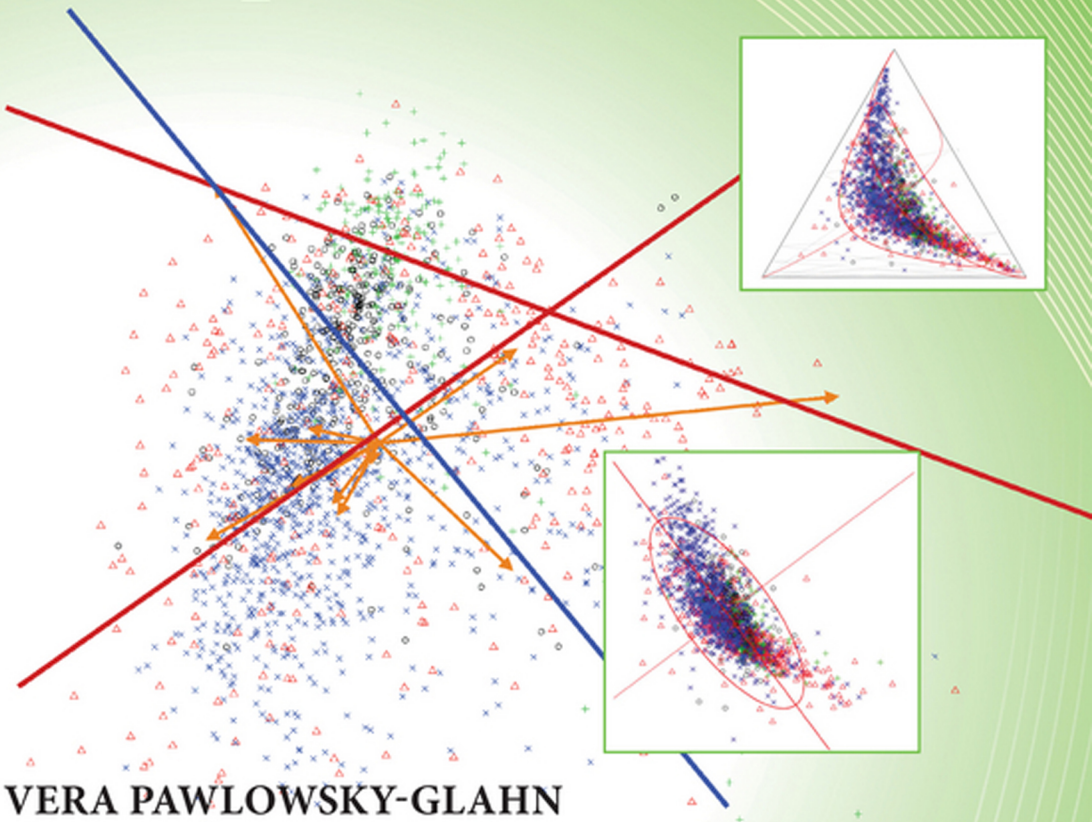
Modeling and Analysis of Compositional Data



VERA PAWLOWSKY-GLAHN
JUAN JOSÉ EGOZCUE
RAIMON TOLOSANA-DELGADO

STATISTICS IN PRACTICE

WILEY

Modeling and Analysis of Compositional Data

STATISTICS IN PRACTICE

Series Advisors

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.

Modeling and Analysis of Compositional Data

Vera Pawlowsky-Glahn

University of Girona, Spain

Juan José Egozcue

Technical University of Catalonia, Spain

Raimon Tolosana-Delgado

Helmholtz Institut Freiberg for Ressources Technology, Germany

WILEY

This edition first published 2015

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Pawlowsky-Glahn, Vera.
Modelling and analysis of compositional data / Vera Pawlowsky-Glahn, Juan José Egozcue, Raimon Tolosana-Delgado.
pages cm
Includes bibliographical references and indexes.
ISBN 978-1-118-44306-4 (cloth)
1. Multivariate analysis. 2. Mathematical statistics. 3. Geometric analysis. I. Egozcue, Juan José, 1950- II. Tolosana-Delgado, Raimon. III. Title.
QA278.P39 2015
519.5′35–dc23

2014043243

A catalogue record for this book is available from the British Library.

ISBN: 9781118443064

Set in 10.5/12.5pt, Times-Roman by Laserwords Private Limited, Chennai, India

1 2015

We cannot solve our problems
with the same thinking
we used when we created them.

Albert Einstein

Eppur si muove

Galileo Galilei

Contents

Preface

This book is an illustration of the adage collected by Thomas Fuller in *Gnomologia* (1732, Adage 560): *All things are difficult, before they are easy* and cited by John Aitchison (1986, Chapter 3). It has been a long way to arrive at this point, and there is still a long and not always easy way to go in the light of the insights presented here. Therefore, we dedicate this work to all those researchers who are not mainstream and have to struggle swimming against the tide.

These pages are based on lecture notes originally prepared as support to a short course on compositional data analysis. The first version of the notes dates back to the year 2000. Their aim was to transmit the basic concepts and skills for simple applications, thus setting the premises for more advanced projects. The notes were updated over the years, reflecting the evolution of our knowledge about the geometry of the sample space of compositional data. The recognition of the role of the sample space and its algebraic-geometric structure has been essential in this process. This book reflects the state of the art at the beginning of the year 2014. Its aim is still to introduce the reader into the basic concepts underlying compositional data analysis, but it goes far beyond an introductory text, as it includes advanced geometrical and statistical modeling. One should also be aware that the theory presented here is a field of active research. Therefore, the learning process can just start with this book, and a study of the latest contributions presented at meetings and as articles in journals is strongly recommended.

The book relies heavily on the monograph "*The Statistical Analysis of Compositional Data*" by John Aitchison (1986) and on posterior fundamental developments that complement the theory developed there, mainly those by Aitchison (1997), Barceló-Vidal et al. (2001), Billheimer et al. (2001), Pawlowsky-Glahn and Egozcue (2001, 2002), Aitchison et al. (2002), Egozcue et al. (2003), Pawlowsky-Glahn (2003), Egozcue and Pawlowsky-Glahn (2005), and Mateu-Figueras et al. (2011). Specific literature for other aspects of compositional analysis is given in the corresponding chapters. Chapter 1 gives a brief overview of the history of these developments and presents some everyday examples to illustrate the need of compositional data analysis. Chapter 2 defines compositions and their characteristics and introduces their sample space,

the simplex. Zeros and other irregular components are addressed in Section 2.3. On the basis of these considerations, Chapter 3 presents the Aitchison geometry of the simplex, while Chapter 4 gathers several ways to represent compositional data within this geometry. These four chapters form the algebraic-geometric body of the book, the backbone of the rest of the material.

Chapter 5 deals with exploratory analysis techniques adapted to compositions. Chapter 6 covers some distribution models for random compositions, as well as some required elements of probability theory. In particular, the latter chapter includes the normal distribution on the simplex, essential for the following two chapters. They are devoted to advanced statistical modeling: Chapter 7 provides some tools for testing compositional hypotheses (numerically and graphically), while Chapter 8 focuses on linear models, including regression, analysis of variance, and discriminant analysis. The last two chapters give an overview of what lies beyond this book: Chapter 9 outlines several compositional models besides the linear model, while the epilogue (Chapter 10) summarizes the ongoing and open aspects of research, as well as further topics, too specific to deserve longer attention in a general-purpose book.

Readers should take into account that, for a thorough understanding of compositional data analysis, a good knowledge in standard univariate statistics, basic linear algebra, and calculus, complemented with an introduction to applied multivariate statistical analysis, is a must. The specific subjects of interest in multivariate statistics, developed under the assumptions that the sample space is the real space with the usual Euclidean geometry, can be learned in parallel from standard textbooks, for instance, Krzanowski (1988) and Krzanowski and Marriott (1994) (in English), Fahrmeir and Hamerle (1984) (in German), or Peña (2002) (in Spanish). Thus, the intended audience goes from advanced students in applied sciences to practitioners, although the original lecture notes proved to be useful for statisticians and mathematicians as well. Newcomers to the field may find specially useful to start with Chapters 1–3, then read the first five sections of Chapter 4 and switch to Chapters 5 and 7 before finishing up Chapter 4. Applied practitioners already familiar with the basics of compositional data analysis should have a look at the notation and concepts in Chapters 4 and 6, before passing to the modeling Chapters 7–9. This book includes an extensive list of references, two appendices with practical recipes and some basic elements of random variables, a list of the symbols used in the book, and two indices; an author index and a general index. In the latter, pages in boldface indicate the point where the corresponding concept is defined.

Concerning notation, it is important to note that, to conform to the standard praxis of registering multivariate observations as a matrix where each row is an observation or data point and each column is a variate, vectors will be considered as row vectors (denoted by square brackets) to make the transfer from

theoretical concepts to practical computations easier. Furthermore, as a general rule, theoretical parameters will be denoted by either Latin or Greek letters and their estimators by the same letters with a hat.

Throughout the book, examples are introduced to illustrate the concepts presented. The end of each example is indicated with a diamond suit ($\diamond$).

Most chapters end with a list of exercises. They are formulated in such a way that many can be solved using an appropriate software. CoDaPack is a user friendly, cross-platform, freeware to facilitate this task, which can be downloaded from the web. Details about this package can be found in Thió-Henestrosa and Martín-Fernández (2005) or Thió-Henestrosa et al. (2005). Those interested in working with `R` (or S-plus) may use the packages "compositions" by Boogaart and Tolosana-Delgado (2005, 2013) in general or "robCompositions" by Templ et al. (2011) for robust compositional data analysis, as well as their common graphical user interface "compositionsGUI" by Eichler et al. (2013).

Vera Pawlowsky-Glahn
Juan José Egozcue
Raimon Tolosana-Delgado

About the Authors

Dr. Vera Pawlowsky-Glahn is professor at the University of Girona, Department of Computer Science, Applied Mathematics, and Statistics. She studied Mathematics at the University of Barcelona (UB), Spain, and obtained her PhD (doctor rerum naturam) from the Free University of Berlin, Germany. Before going to Girona, she was professor in the School of Civil Engineering at the Technical University of Catalonia (UPC) in Barcelona. Her main research topic since 1982 has been the statistical analysis of compositional data. The results obtained over the years have been published in multiple articles, proceedings, and books. Together with A. Buccianti she has acted as editor of a book in honor of J. Aitchison in 2011 published by Wiley, who will also publish in 2015 a textbook on modeling and analysis of compositional data, co-authored with J. J. Egozcue and R. Tolosana-Delgado. She was the leader of a research group on this topic involving professors from different Spanish universities. The group organizes every two years a workshop on compositional data analysis, known as CoDaWork, and their research has received regularly financial support from the Spanish Ministry for Education and Science and from the University Department of the Catalan Government. Prof. Dr. Pawlowsky-Glahn has been vice-chancellor at UPC from 1990 to 1994, head of the Department of Computer Science and Applied Mathematics at the University of Girona in 2004–2005, and dean of the Graduate School of the University of Girona in 2005–2006. She received in 2006 the William Christian Krumbein Medal of IAMG, was nominated Distinguished Lecturer of IAMG in 2007, and received the J.C. Griffiths Teaching Award in 2008. From 2008 to 2012 she was President of IAMG and is now Past-President.

Dr. Juan José Egozcue studied Physics, oriented to Geophysics and Meteorology, at the University of Barcelona (Spain). He obtained his PhD in Physics in the same university with a dissertation on maximum entropy spectral analysis (1982). In 1978 he got a position as a lecturer in the school of civil engineering in Barcelona (Escuela de Ingeniería de Caminos, Canales y Puertos de la Universidad Politécnica de Cataluña (UPC), Barcelona, Spain), teaching several topics on Applied Mathematics. In 1983 he started teaching Probability and Statistics. He became Full Professor in 1989, at the UPC, where he has been Vice-Chancellor of the university (1986–1988) and Chair of the Department of Applied Mathematics III (1992–1998).

His research activities are presently centered in two lines: estimation of natural hazards using Bayesian methods, specially applied to seismic, rainfall and ocean wave hazards; and analysis of compositional data, with special emphasis in the geometry of the sample space.

He started research on compositional data analysis around 2000 in cooperation with Dr. Vera Pawlowsky-Glahn. The first results appeared in 2001–2002 when the Euclidean vector space structure of the simplex recognized. The development of this geometry led to the introduction of the isometric logratio transformation for compositional data and the concept of balance (2003–2005) which have proven their usefulness in a number of applied fields.

Dr. Raimon Tolosana-Delgado completed in 2002 a degree in Engineering Geology in Barcelona, at the School of Civil Engineers (UPC) and the Faculty of Geology (UB), and in 2004 a Master in Environmental Technology and Physics at UdG, all along focusing on compositional data analysis and Geostatistical methods applied to Earth Sciences. In 2006 he completed his PhD under the supervision of Dr. Pawlowsky-Glahn, on the spatial analysis of data from restricted spaces, as a generalization of compositional data analysis. Since then, he has been working as a fellow researcher between Spain and Germany, working with compositional models in sedimentology at the University of Göttingen, and later back at the UPC, in weather and climate modeling and data assimilation through geostatistical simulations and restricted space consideration. Since October 2012 he has been applying and developing

compositional and spatial methods as a researcher at the Department of Modeling and Valuation, Helmholtz Institute Freiberg for Resource Technology in Freiberg (Saxony, Germany), a joint research institute of the Technical University "Bergakademie" Freiberg and the Helmholtz Zentrum Dresden-Rossendorf dealing with all aspects of the value chain of Rare Earths and other technological elements, from mineral exploration and mining to technological waste recycling.

Acknowledgments

We acknowledge the many comments made by readers of the original lecture notes, pointing at both small and important errors in the text. Essential have also been the many contributions and discussions presented at several editions of CoDaWork, the International Workshop on Compositional Data Analysis, extensively cited throughout the text. They all have contributed to improve the theory presented here. We also appreciate the support received from our institutions, research groups, the *Spanish Ministry of Economy and Competitiveness* under the project "METRICS" (Ref. MTM2012-33236), and the Generalitat de Catalunya through the project "Compositional and Spatial Analysis" (COSDA) (Ref. 2014SGR551).

1

Introduction

Compositional data describe parts of some whole. They are commonly presented as vectors of proportions, percentages, concentrations, or frequencies. As proportions are expressed as real numbers, one is tempted to interpret, or even analyze, them as real multivariate data. This practice can lead to paradoxes and/or misinterpretations, some of them well known even a century ago, but mostly forgotten and neglected over the years. Some simple examples illustrate the anomalous behavior of proportions when analyzed without taking into account the special characteristics of compositional data.

Example 1.1 (Intervals covering negative proportions).
Daily measurements of an air pollutant are reported as $3 \pm 5\ \mu g/m^3$. The given interval of concentration covers a nonsensical range of concentrations that includes negative values. It is probably generated by an average of concentrations which contain some values much higher than $3\ \mu g/m^3$. For instance, the following is a set of rounded random percentages: $1, 1, 2, 3, 4, 4, 7, 13, 29, 37$. Their mean is 10.1%, while their standard deviation is 12.7%. Thus a typical $2s$-interval for the mean value would be an interval covering negative proportions, namely, $(-15.3\%; 35.5\%)$. A frequent procedure is to cut this interval at zero, but then the question arises on what happens to the probability assigned to the eliminated part of the interval, $(-15.3\%; 0\%)$, and to the probability assigned to the retained part, $(0\%; 35.5\%)$. ◇

Modeling and Analysis of Compositional Data, First Edition.
Vera Pawlowsky-Glahn, Juan José Egozcue and Raimon Tolosana-Delgado.

Example 1.2 (Small proportions: Are they important?).
Frequently, when some components or parts of a composition are very small, they are eliminated, with the argument that they are negligible. In such a case, it is important to think about *the salt in a soup*. Consider a soup that is perfectly seasoned to your taste, and imagine somebody adds to the soup the same amount of salt you used, thinking that it was not yet seasoned. Probably, doubling the amount of salt will spoil it completely. To our understanding, this is a perfect example on how important a small proportion can be and why a relative scale gives you better information in this case than an absolute one. Sometimes, small proportions are added to other parts, for example, salt and other spices, but that leads to a loss of information, making the recipe insufficiently specified. ◇

Example 1.3 (Reporting changes in proportions).
In the 1998 election to the German Bundestag, the German Liberal Party (FDP) obtained 6.2% of the votes. Eleven years later, in the 2009 elections, they obtained a share of 14.6%. This could be reported as an increment of 8.4 percentage points. We are more used to reading that FDP increased its proportion of votes a 135% ($6.2 + 6.2 \times 135/100 = 14.6$). In the following election, just 4 years later, the party decreased its votes by a significant 67%, but still half of the increment that occurred between 1998 and 2009. Nevertheless, that meant that the FDP was not anymore represented in the Bundestag, because its share ($14.6 - 14.6 \times 67/100 = 4.8$) dropped below the threshold of 5% required by the German electoral law. How can it be that increasing 135% and decreasing 67% gives a negative balance? Perhaps this is a bad way of reporting changes in proportions (data extracted from Wikipedia (2014)).

Reporting increments of shares in differences of percentage points have also disappointing properties, as the relative scale of proportions is ignored. In fact, an increment of 8.4 percentage points represents a very important change from the 1998 result of FDP (6.2%). It would be not so important if the previous 1998 result were, for instance, 30%. ◇

Example 1.4 (The scale of proportions).
In a given year, the annual proportion of rainy days in a desert region is 0.1%, and near a mountain range it is 20.0%. Some years later, these proportions have changed to 0.2% and 20.1%, respectively. To summarize the situation, one can assert that the rainy days in both regions have increased by 0.1%. Such a statement suggests the idea of a homogeneous change in the two different regions, ignoring that the rainy days in the desert have been doubled, while in the mountain range the proportion is almost the same. Using the increment of ratios typical of election results or economic reports, the rainy days would have increased a critical 100% in the desert, and a slightly relevant 5% in the mountains.

Furthermore, if some analysis of the evolution of the rainy days is made in both regions, it should be guaranteed that equivalent results are obtained if the nonrainy days are analyzed. In the desert region, the annual proportion of nonrainy days has changed from 99.9% to 99.8% and near the mountain range from 80.0% to 79.9%. That represents that nonrainy days have decreased, respectively, 0.001% and 0.00125%, which suggests almost no difference between the mountain and the desert. How can it then be that rainy days change so dramatically in the desert and nonrainy days do not change at all? A proper analysis should assure that no paradoxical results are obtained when analyzing one type of days and its complementary. ◇

Example 1.5 (The Simpson's paradox).
The lectures on statistics started very early this morning. Students (men and women) are divided into two classrooms. Some of them arrived on time and some of them were late. Academia was interested in knowing about punctuality according to the gender of the students. Therefore, data were collected this morning during the statistics lectures. The data set is reported in Table 1.1. The paradoxical result is that, for both classrooms, the proportion of women arriving on time is greater than that of men. On the contrary, if the individuals of both classrooms are joined in a single population, the proportion of punctual men is larger than that of the women. This kind of paradoxical results are known as Simpson's paradox (Simpson, 1951; Julious and Mullee, 1994; Zee Ma, 2009). The paradox can be viewed from different points of view. The simplest one, the arithmetic perspective, is to look at the way in which proportions are aggregated: to find the proportion of on-time women in the joint population, the per classroom proportions a_1/W_1, a_2/W_2 are *averaged* as $(a_1 + a_2)/(W_1 + W_2)$, where a_i is the number of on-time women in the classroom i and W_i is the corresponding

Table 1.1 Number of students of two classrooms, arriving on time and being late, classified by gender. Proportions are reported under the number of students. The largest proportion of arriving on-time men and women are in boldface for easy comparison.

	Classroom 1		Classroom 2		Total	
	On time	Late	On time	Late	On time	Late
Men	53	9	12	6	65	15
	0.855	0.145	0.667	0.333	**0.813**	0.188
Women	20	2	50	18	70	20
	0.909	0.091	**0.735**	0.265	0.778	0.222

total of women. This kind of average is ill-behaved for proportions as shown by Simpson's paradox.

A second point of view is to look at the total proportion of on-time women as a mean value of this proportion in the two classrooms. Each classroom is treated as a sample individual and $(a_1 + a_2)/(W_1 + W_2)$ is taken as the sample mean of the proportions. The paradoxical result suggests that mean values of proportions should be redefined carefully to get consistent results. ◇

Example 1.6 (Spurious correlation).
The Spanish Government publishes the number of affiliations to the Social Security on a monthly basis, which is classified into the following categories depending on the type of company: agricultural, industrial, construction, and service. The 144 data, corresponding to a monthly series going from 1997 to 2008, were downloaded from the corresponding web site (Gobierno de España, 2014). A version, prepared for processing, is available in (www.wiley.com/go/glahn /practical). First, to obtain proportions between the different types of company, the data were normalized to add to 1 in the full composition comprising the four categories. Then, the correlation matrix was computed (see Table 1.2). Next, to analyze the behavior of the companies excluding *construction*, a subcomposition of three categories was obtained, suppressing the category *construction* and converting the three-part vector to proportions, so that the three components add up to 1. Again, the correlation matrix was computed (see Table 1.3). When analyzing correlations in the full composition with four parts and the subcomposition with three parts, the correlation between the proportion of agricultural and industrial companies only changed slightly, actually from −0.9808 to −0.9887, whereas the correlation between the service companies and either agricultural or industrial companies changed dramatically, from 0.1699 to 0.9863 in the first case and from −0.0723 to −0.9999 in the second. This is a typical effect when analyzing a set of parts adding up to a constant, or a subset of the same parts, closed to any constant.

Table 1.2 Correlation of proportion of affiliations to social security in Spain according to the type of company (four-part composition: agricultural, industrial, construction, and service).

	Agricultural	Industrial	Construction	Service
Agricultural	1.0000	−0.9808	0.9201	0.1699
Industrial	−0.9808	1.0000	−0.9663	−0.0723
Construction	0.9201	−0.9663	1.0000	−0.1867
Service	0.1699	−0.0723	−0.1867	1.0000

Table 1.3 Correlation of proportion of affiliations to social security in Spain according to the type of company (three-part subcomposition: agricultural, industrial, and service).

	Agricultural	Industrial	Service
Agricultural	1.0000	−0.9887	0.9863
Industrial	−0.9887	1.0000	−0.9999
Service	0.9863	−0.9999	1.0000

The problem of spurious correlation is sometimes circumvented by avoiding the closure when considering a subcomposition. This is equivalent to say: the percentages of agricultural, industrial, construction, and service affiliates constitute a composition as the percentages add to 100%; to overcome the compositional intricacies, we can remove one component, for example, service, so that the remaining percentages do not add to 100%. This way, the correlation matrix between the percentages of agricultural, industrial, and construction affiliates are exactly those reported in Table 1.2 in the first three columns and rows. However, a new question arises: what would happen if we start with two additional categories of affiliation closed to 100%? ◇

The awareness of problems related to the statistical analysis of compositional data dates back to a paper by Karl Pearson (1897) the title of which began significantly with the words *"On a form of spurious correlation ... "*. Since then, as stated in Aitchison and Egozcue (2005), the way to deal with this type of data has gone through roughly four phases, which can be summarized as follows:

Phase I: 1897–1960

Karl Pearson, in his paper on spurious correlations, pointed out the problems arising from the use of standard statistical methods with proportions. But his warnings were ignored until around 1960, despite the fact that a compositional vector – with components the parts of some whole – is usually subject to a constant-sum constraint.

Phase II: 1960–1980

Around 1960, the geologist Felix Chayes (1960) took up the problem and warned against the application of standard multivariate analysis to compositional data. He tried to separate what he called the *real* from the *spurious* correlation, in an attempt to avoid the *closure problem*, expressed mainly as a negative bias induced by the constant-sum constraint. Important contributions in geological

applications were made, among others, by Sarmanov and Vistelius (1959), and Mosimann (1962) which drew the attention of biologists. However, as pointed out by Aitchison and Egozcue (2005), *distortion of standard multivariate techniques when applied to compositional data was the main goal of study*.

Phase III: 1980–2000

Aitchison, in the 1980s, realized that compositions provide information about relative, not absolute, values of parts or components. Consequently, every statement about a composition can be stated in terms of ratios of components (Aitchison, 1981, 1982, 1983, 1984). The facts that logratios are easier to handle mathematically than ratios, and that a logratio transformation provides a one-to-one mapping onto a real space, led to the advocacy of a methodology based on a variety of logratio transformations. These transformations allowed the use of standard unconstrained multivariate statistics applied to transformed data, with inferences translatable back into compositional statements. But they were, not without difficulties, derived from the fact that the usual Euclidean geometry and measure were implicitly assumed for the sample space of compositional data.

This phase deserves special attention because transform techniques have been very popular and successful over more than a century; from the Galton-McAlister introduction of the logarithmic transformation for positive data, through variance-stabilizing transformations for sound analysis of variance, to the general Box–Cox transformation (Box and Cox, 1964) and the implied transformations in generalized linear modeling. The logratio transformation principle is based on the fact that there is a one-to-one correspondence between compositional vectors and associated logratio vectors, so that any statement about compositions can be reformulated in terms of logratios, and vice versa. The advantage is that the problem of a constrained sample space, the simplex, is removed. Data are projected into multivariate real space, opening up all available standard multivariate techniques. The original transformations were principally the additive logratio transformation (Aitchison, 1986, p. 113) and the centered logratio transformation (Aitchison, 1986, p. 79). The logratio transformation methodology seemed to be accepted by the statistical community; see, for example, the discussion of Aitchison (1982).

Phase IV: 2000–present

Around 2000, several scientists realized independently that the internal simplicial operation of perturbation, the external operation of powering, and the simplicial metric define a metric vector space (indeed a Hilbert space) (Billheimer et al., 1997, 2001, Pawlowsky-Glahn and Egozcue, 2001). The recognition of